[WEDT.A] Dokumentacja końcowa projektu
Klasyfikacja stron WWW na podstawie struktury

Michał Aniserowicz <michalaniserowicz@gmail.com>
Jakub Turek jkbturek@gmail.com

30 maja 2013r.

1 Temat projektu

Tematem projektu jest automatyczna klasyfikacja stron WWW na podstawie struktury. W ramach uścieślenia tematu projektu, wybrane do rozpoznowania zostały następujące kategorie stron internetowych:

- dzienniki internetowe (*blogi*),
- strony społecznościowe oparte na obrazkach (*kwejki*),
- serwisy informacyjne,
- sklepy internetowe.

Projekt został wykonany w technologii Python 2.7.4 i był testowany na systemach operacyjnych Windows 7 oraz Ubuntu 13.04. Projekt wykorzystuje bibliotekę PIL[1] w wersji 1.1.7.

2 Implementacja

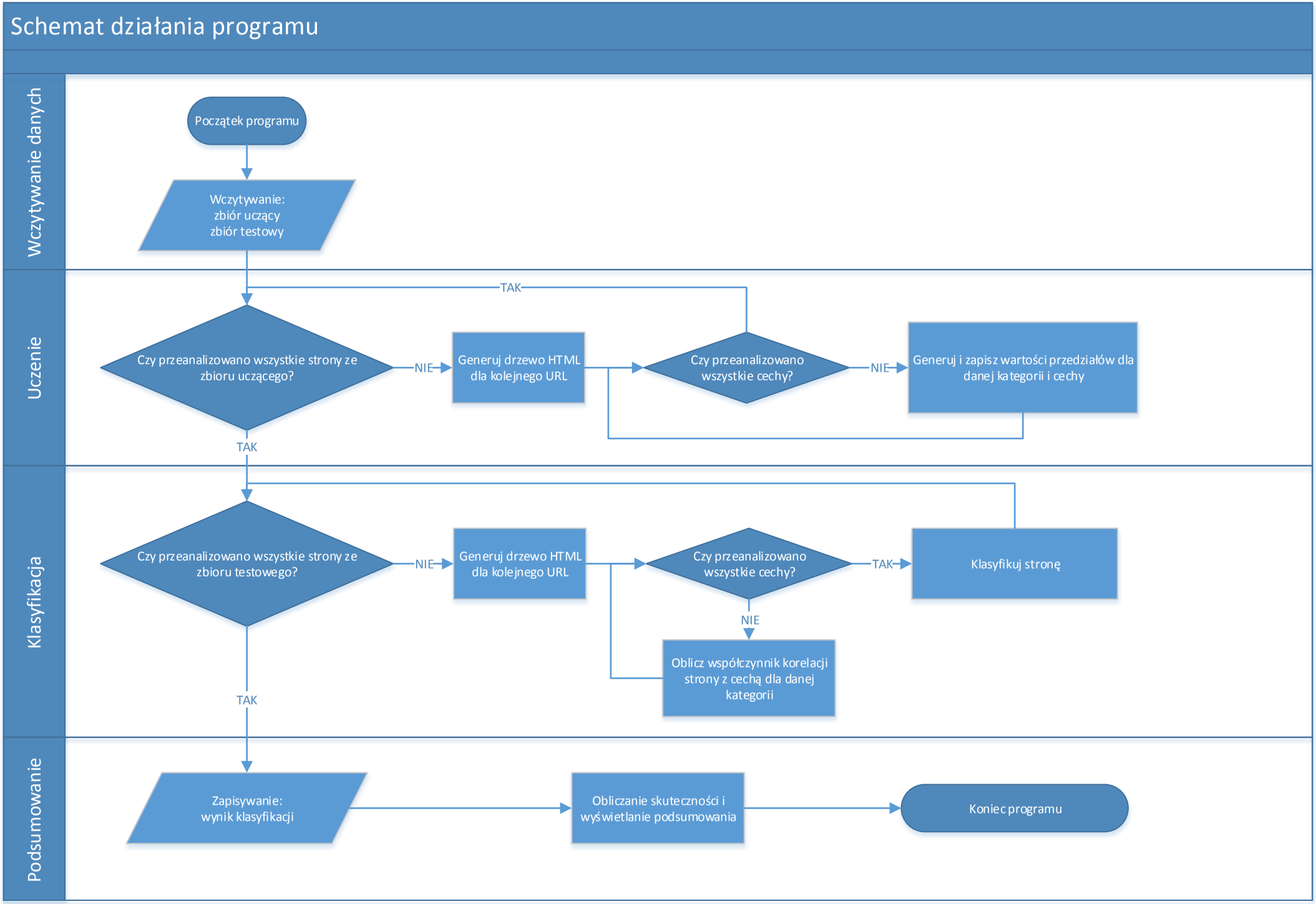
2.1 Schemat działania aplikacji

Na następnej stronie przedstawiony został ogólny schemat działania programu. Obejmuje on dwie główne fazy działania aplikacji:

uczenie się Program generuje zestawienie wartości cech dla poszczególnych kategorii na podstawie próby wzorców.

klasyfikacja Program dokonuje klasyfikacji pozostałych stron na podstawie wartości cech wyznaczonych w poprzednim kroku.

[1] Python Image Library - http://www.pythonware.com/products/pil/.

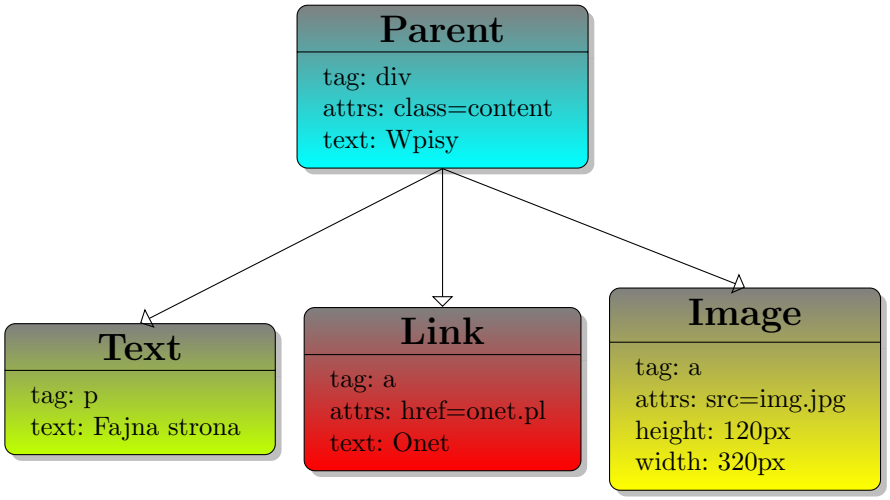

Schemat działania programu
Wczytywanie danych
Uczenie
Klasyfikacja
Podsumowanie
Początek programu
Wczytywanie:
zbiór uczący
zbiór testowy
Czy przeanalizowano wszystkie strony ze zbioru uczącego?
NIE
Generuj drzewo HTML dla kolejnego URL
Czy przeanalizowano wszystkie cechy?
NIE
Generuj i zapisz wartości przedziałów dla danej kategorii i cechy
TAK
TAK
Czy przeanalizowano wszystkie strony ze zbioru testowego?
NIE
Generuj drzewo HTML dla kolejnego URL
Czy przeanalizowano wszystkie cechy?
TAK
Klasyfikuj stronę
NIE
Oblicz współczynnik korelacji strony z cechą dla danej kategorii
TAK
Zapisywanie:
wynik klasyfikacji
Obliczanie skuteczności i wyświetlanie podsumowania
Koniec programu

2.2 Drzewo HTML

Strony WWW są wewnętrznie reprezentowane przez drzewo HTML. Każdy korzeń drzewa posiada następujące atrybuty:

- tag,
- słownik atrybutów (np. `class="main-img", src="image.jpg"`),
- tekst wewnątrz taga,
- wysokość elementu,
- szerokość elementu.

Ponadto od dowolnego korzenia można dojść do jego rodzica, a także wszystkich jego dzieci. Rysunek 1 przedstawia schemat wykorzystywanego drzewa HTML.



Rysunek 1: Poglądowy schemat drzewa HTML.

Etap budowy drzewa jest w pełni konfigurowalny, dzięki użyciu następujących atrybutów:

dozwolone tagi Lista tagów, z których mogą powstawać liście drzewa. Jeżeli tag nie znajduje się na dozwolonej liście, tekst znajdujacy się w jego środku jest konkatenowany z tekstem jego pierwszego dozwolonego rodzica.

dozwolone atrybuty Lista atrybutów, które są włączane do słownika w liściach drzewa. Pozostałe atrybuty i ich wartości są pomijane.

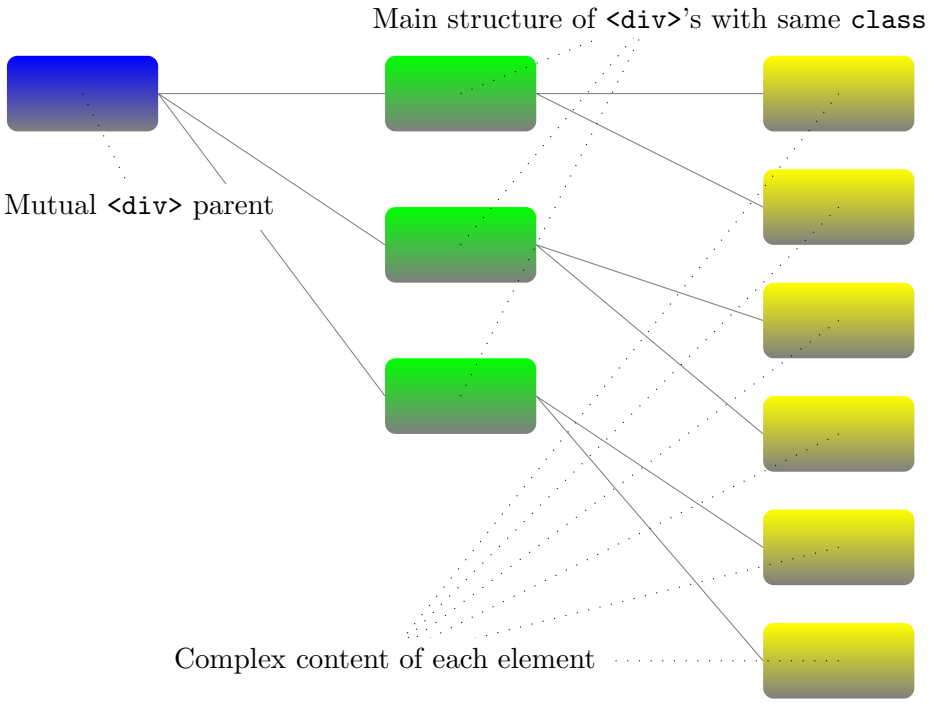
zakazane tagi Lista tagów, które zawierają tekst nie włączany do pierwszego dozwolonego rodzica. Umożliwia to odfiltrowywanie m.in. skryptów.

2.3 Główna struktura strony

Przed omówieniem zestawu cech badanych przez aplikację, należy wprowadzić pojęcie *głównej struktury strony*, która będzie intensywnie analizowana. Główna struktura strony to najliczniejsza struktura, która posiada następujące cechy:

- Występują w niej wyłącznie tagi `<td>` lub `<div>`, przy czym w danej strukturze są to tagi wyłącznie jednego z tych typów.
- Wszystkie tagi występują na jednym poziomie głębokości w drzewie. Oznacza to, że mają wspólnego rodzica.
- Każdy z tagów posiada jednakową wartość atrybutu `class`.
- Każdy element struktury posiada więcej niż jeden element podrzędny lub większy niż jeden poziom zagłębienia elementów.

Główna struktura jest zaprezentowana na rysunku 2.



Rysunek 2: Zielone elementy należą do głównej struktury strony.

Struktura ta jest charakterystyczna dla wszystkich czterech typów klasyfikowanych stron. Na *blogach* zawiera treść wpisów, na *kwejkach* obrazki, na stronach informacyjnych odnośniki do artykułów, natomiast w sklepach internetowych - odnośniki do kategorii i/lub produktów.

2.4 Analizowane cechy strony

Na etapie uczenia się oraz klasyfikacji brane są pod uwagę, między innymi, następujące cechy:

- Liczba elementów w głównej strukturze strony.
- Liczba powtórzeń głównej struktury strony. Sprawdza czy struktura się powtarza i, jeżeli tak, to w jakiej liczbie.
- Średnia ilość tekstu przypadająca na każdy element głównej struktury strony oraz jego dzieci.
- Średnia ilość obrazów przypadająca na każdy element głównej struktury strony oraz jego dzieci.
- Liczba obrazów na stronie.
- Rozmiary największego oraz najmniejszego obrazu na stronie.
- Liczba odnośników na stronie.
- Stosunek długości tekstu zawartego w odnośnikach do długości całego tekstu zawartego na stronie.
- Średnia długość lokalnego odnośnika[2] na stronie.
- Liczba tagów w specyfikacji HTML5 (przykładowo: `<article>` oraz `<section>`).
- Stosunek liczby odnośników do liczby obrazów na stronie.
- Stosunek długości tekstu do liczby obrazów na stronie.

Cechy te dobrze różnicują cztery zadane kategorie stron internetowych. Przykładowo serwisy informacyjne charakteryzują się wysokim współczynnikiem długości tekstu w odnośnikach do całkowitej długości tekstów, długimi odnośnikami lokalnymi, dużą ilością obrazów na stronie oraz częstym występowaniem tagów w specyfikacji HTML5. Dla kontrastu blogi charakteryzują się niewielkim stosunkiem liczby odnośników do długości tekstu, niewielką liczbą obrazków, dużą koncentracją tekstu w głównej strukturze oraz większą niż dla serwisów informacyjnych liczbą elementów w głównej strukturze strony.

[2]Odnośnik lokalny prowadzi do tej samej domeny, w której znajduje się analizowana strona.